Population structure, demography and admixture

G. Hellenthal

University College London Genetics Institute (UGI), Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

The increasing availability of large-scale genetic variation data sampled from world-wide geographic areas, coupled with advances in the statistical methodology to analyse these data, is showcasing the power of DNA as a major tool to gain insights into the history of humans and other organisms. This chapter describes the concepts behind some widely-used methods in the field of population genetics applied to whole genome autosomal data. In particular, the chapter will focus on techniques that analyse genetic data in order to learn about sub-structure among sampled individuals and the dynamics of population size changes and intermixing among genetically different populations. While this is by no means an exhaustive look at the many interesting methods in the field, it provides an overview of some of the demographic signals inherent in genetic data and the challenges in extracting this information. We will describe these methods as if they will be applied to data from humans, though note the concepts here extend to other diploid organisms that experience homologous recombination (with potentially straight-forward extensions to organisms of other ploidy as well). In general this chapter illustrates the power of DNA as an important data resource that, when interpreted in the context of knowledge from other data sources (e.g. archaeology, anthropology, linguistics), can resolve existing controversies or unearth previously unknown features of human history.

1 Introduction

The ancestral history of anatomically modern human groups is exceedingly complex. Demographic factors affecting genetic variation data include *population splits* where different groups ("populations") of individuals become isolated from one another, after which each group is subjected to independent genetic drift that results in allele frequency differences between them. In addition, changes in the sizes within each population, in terms of the effective number of breeding individuals, can alter the speed at which drift acts, with population expansions and contractions (e.g. "bottlenecks") decreasing and accelerating the effects of drift, respectively. Another important process is *admixture*, where previously isolated groups intermix. As discussed below, even the concept of a "population" is not straight-forward, as individuals can be grouped together in many different ways.

Each of these processes affects genetic variation patterns in distinct ways, yet disentangling all of the possible processes that can lead to observed variation patterns is an intractable statistical problem. Nonetheless advances have been made to attempt to distinguish the effects of some of these features on DNA, taking advantage of the vast amounts of genetic information currently available. A key insight is that genetic patterns are correlated among sequences from different individuals, including among individuals that are unrelated at the familial level (i.e. individuals that are not first or second cousins, etc), with such unrelated individuals being the focus of this chapter. This correlation carries vital information on the extent of shared recent ancestry among such samples of unrelated individuals. Therefore by studying genetic correlations, we can hope to learn about the degree to which different sets of individuals, e.g. sampled from different geographic regions, are ancestrally related to one another. While every pair of individuals shares a common ancestor at each point of the genome, approaches here attempt to determine which individuals share ancestors that lived more recently than the shared ancestors of other individuals.

This chapter will concentrate on four specific types of inference:

1. exploring spatial summaries of genetic variation data
2. classifying individuals into clusters based on genetics
3. inferring population size changes and split times
4. identifying and describing admixture events

This chapter will not exhaustively explore all methods related to (1)-(4), but instead will provide insights into some commonly-used approaches applicable to genome-wide autosomal data that address these questions. We will discuss the patterns in genetic variation data that theoretically allow inference under each approach, providing an overview of some of the mathematical details. We will also highlight some applications and limitations of each.

1.1 "Admixture" versus "background" linkage disequilibrium

As described in more detail elsewhere (McVean chapter), the non-random association among allelic types at different genetic loci is called *linkage disequilibrium (LD)*. As this is a principal feature used for inference in many of the approaches discussed in this chapter, it is helpful to define different types of LD based on the factor driving the association. Consider the admixed population formed as illustrated in Figure 1, which shows a single "pulse" of admixture (i.e. admixture occurring over a short time interval, such as one generation) between two populations. Subsequently, individuals in the newly admixed population mate randomly for r generations. As shown at the bottom of Figure 1, DNA in these admixed individuals will be mixtures of segments ("tracts") inherited intact from individuals from the admixing source populations. Therefore loci within the same block can be correlated due to inheritance from a common recent ancestor from this admixture event. Following [1], we will refer to this as *admixture LD*. As discussed below, depending on the date of admixture r, admixture LD can extend over relatively large segments of an autosome, e.g. over megabases.

A second, distinct type of LD we will refer to as *background LD*, again following the terminology of [1], which measures the rate of decay of associations among loci *within* a population, e.g. within each solid-black and dashed-line bar of Figure 1. The level of background LD may be different within each admixing population, and reflects a population's demographic history, including e.g. previous admixture, population size changes and population substructure. As an example, if one of the admixing populations experienced a strong bottleneck (e.g. due to a founder event) and the others did not, we would usually expect its background LD to extend farther than that of the other populations. In general, as the physical distance between two loci increases, the level of background LD between the loci decays at a much faster rate relative to the level of admixture LD between them, with background LD decaying on the order of tens to hundreds of kilobases in humans [2]. This is because background LD captures the effects of demographic processes occurring over the entire history of a population, which can span a very large time frame and hence is affected by many historical recombination events, while admixture LD captures only the effects of more recent specific admixture events. A consequence of its faster rate of decay is that background LD is only potentially significant among loci that are physically quite close to each other. While some approaches described below ignore background LD or attempt to remove it, this chapter also highlights methods that attempt to exploit background LD to increase precision when inferring demographic parameters.

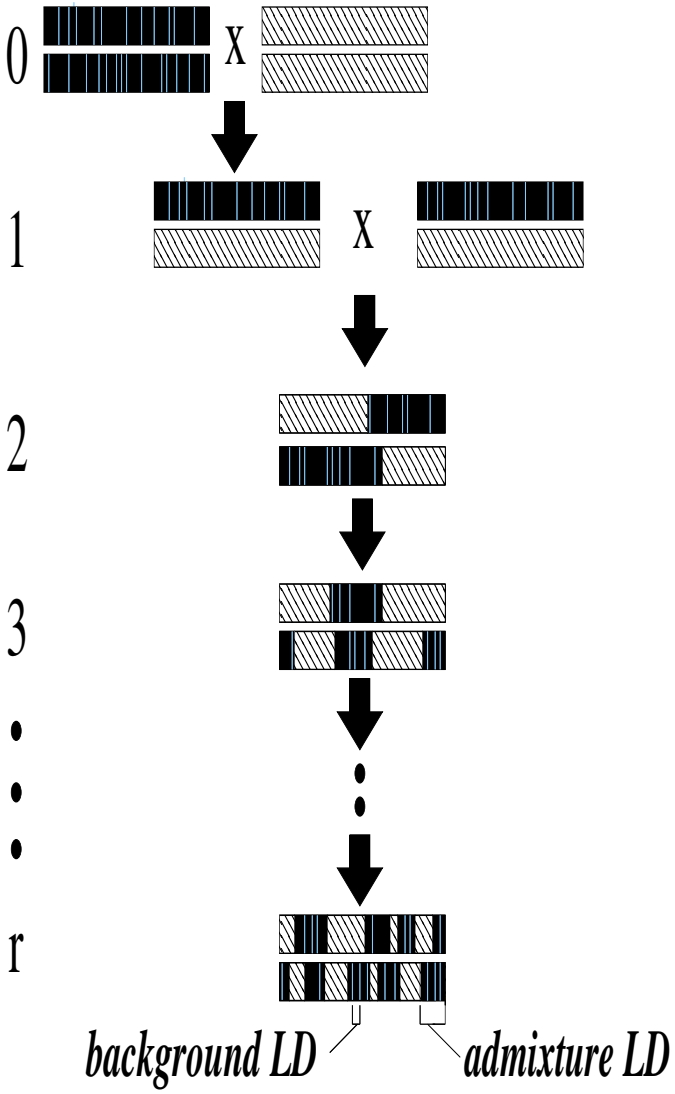


Figure 1: Schematic of the effects on an autosomal region of admixture occurring r generations ago between two populations. Two genetically distinct populations, with DNA represented by solid black and dashed lines, admix at generation 0 at top. For simplicity, the two chromosomes for one individual per population is shown at top. This admixture event is followed by r generations of subsequent random mating among individuals from the admixed population. (For simplicity, after generation 1 we show the two chromosomes of one individual each generation for this autosomal region.) In the first generation following admixture (second row from top), in this genetic region each admixed individual receives a chromosome from an individual representing each population. Subsequent generations inherit "tracts" of contiguous DNA segments from each admixing source, with the lengths of these segments getting smaller each generation due to recombination. As an illustration of "background LD", light blue lines within each solid black bar reflect boundaries of tracts inherited from ancestors living at a time $\gg r$. Loci within each of these tracts will be in "background LD", while loci inherited within the same contiguously solid black (or dashed line) block at bottom will be in "admixture LD".

2 Spatial summaries of genetic variation using principal-components-analysis

A widely-used technique to visualize genetic patterns in a dataset applies principal components analysis (PCA) to the high-dimensional genetic variation data of sampled individuals, and then plots low-dimensional *principal components* that efficiently summarize these data. First proposed in the context of analysing genetic variation data by Cavalli-Sforza and colleagues [3], PCA is an algebraic technique for summarizing variation in multi-dimensional datasets in order to potentially highlight interesting patterns.

Assuming each locus is biallelic in the sample, such as Single Nucelotide Polymorphism (SNP) data, let $X_i \equiv \{X_{i1}, ..., X_{iL}\}$ be the genotype for individual $i \in [1, ..., N]$ at all L sampled biallelic loci, with X_{il} the genotype at locus l of individual i. Specifically, X_{il} will be 0, 1, or 2, reflecting the number of copies of a particular allele that diploid individual i carries

at locus l. Let X represent the $L \times N$ matrix with columns equal to X_i. Typically the matrix X is standardised in some manner, with different standarisations proposed by e.g. [4] and [5], creating a new $L \times N$ matrix Y that e.g. subtracts the row average so that

$$Y_{il} = X_{il} - (1/N) \sum_{i=1}^{N} X_{il},$$

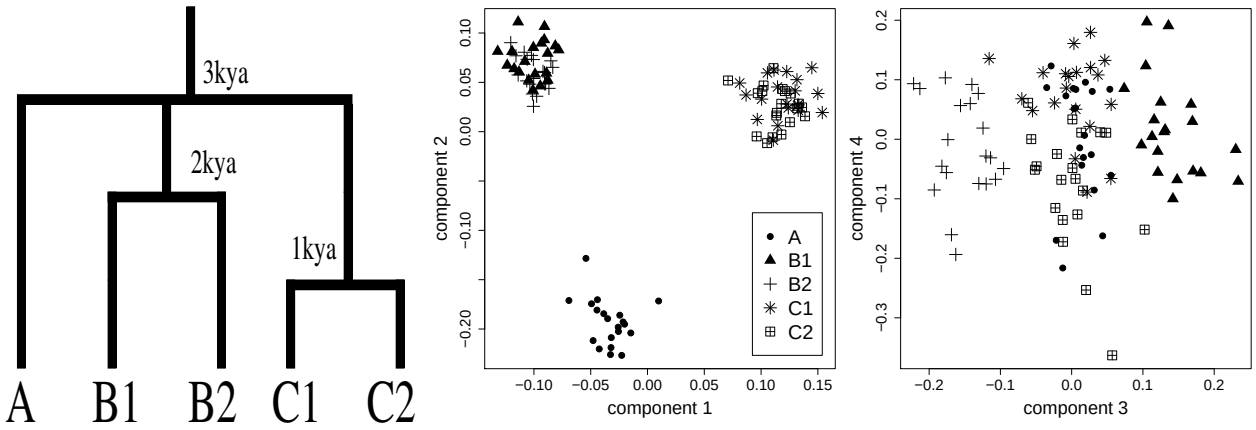
or subtracts the row average and standardizes by the variance so that

$$Y_{il} = \left[X_{il} - (1/N) \sum_{i=1}^{N} X_{il}\right] / \sqrt{v_l(1 - v_l)},$$

where v_l estimates the allele frequency of locus l. (For example, [4] use $v_l = (1 + \sum_{i=1}^{N} X_{il})/(2 + 2N)$.) Relative to the former, this latter standardisation upweights the relative influence on the PCA from variants with low minor allele frequencies among the sampled individuals. Then the N *eigenvectors* and *eigenvalues* of the $N \times N$ matrix $\Omega = Y^T Y$ can be determined using e.g. singular value decomposition (e.g. [5, 6]). Each eigenvector contains N values, with the ith value a linear combination of the genotype data at all L loci of individual i, i.e. a summary of all data points for individual i. An attractive property of PCA is that the eigenvectors are mutually orthogonal, so that they each capture independent information in the data. Furthermore, the first eigenvector explains more overall variation in the data Y than the second eigenvector, and so forth. Therefore, plotting the first two eigenvectors can summarise the strongest signals in the genetic variation data when using only two datapoints per individual, rather than using the L datapoints that contain the full genetic information.

As an example, we applied the PCA software EIGENSTRAT [4] to simulated data from [7]. These simulations consist of genetic variation data from five simulated populations that are related according to the tree in Figure 2a, i.e. where populations A, B={B1,B2} and C={C1,C2} split from each other 3,000 years ago (3kya), followed by sub-populations B1 and B2 splitting from each other 2kya and sub-populations C1 and C2 splitting from each other 1kya. Twenty individuals were sampled from each population {A,B1,B2,C1,C2}, each having SNP data for 150 five megabase regions simulated to mimic sequencing data, with ≈24K SNPs per region after removing singletons (see [7] for more details). As can be seen in the PCA results of Figure 2b, the first two eigenvectors cleanly separate populations A, B and C, suggesting that the strongest signal in these data are the genetic differences between these three groupings. Eigenvector 3 further separates sub-populations B1 and B2, while eigenvector 4 partially separates (though with substantial overlap) sub-populations C1 and C2. In total there are N eigenvectors that can be plotted, though each explains progressively less variation in the total data. Typically the first few eigenvectors are the most informative about underlying population sub-structure, for example separating different populations within continental regions such as Africa, Asia [5] and Europe [8, 9].

While PCA is a widely-used means of summarising large-scale genetic data, interpreting the past demographic processes leading to PCA patterns can be challenging. McVean [10] showed how the expectation of Ω can be related to the mean time of coalesence between pairs of samples. However, PCA projection can depend strongly on sample size and the ascertainment of samples, and different ancestral histories can lead to very similar PCA patterns [10, 11]. Related techniques such as multidimensional scaling have also been employed on genetic data [12], which have similar advantages and disadvantages to PCA.



(a) simulated populations (b) PCA of simulated data

Figure 2: (a) Graphical depiction of tree relating simulated populations {A,B1,B2,C1,C2} from [7]. (b) First four principal components of a PCA applied to the genotype data for the simulations of populations {A,B1,B2,C1,C2}, as calculated by EIGENSTRAT, with each point a simulated individual from one of the 5 populations.

3 Clustering algorithms

3.1 Defining "populations"

Another widely-used means of summarising genetic variation data is to cluster individuals based on associations in their genetic patterns. To some extent, clustering can provide insights into the processes leading to observed genetic patterns. Another motivation for clustering is that researchers often wish to study the demographic processes that have affected a particular "population". While many studies define populations using self-described labels from the individuals sampled or using the geographic sampling location of these individuals, Figure 2b suggests that there is clear utility for classifying individuals into discrete groups using solely genetic variation data. Indeed this classification may be particularly useful when exploring population demography, as individuals who identify with a particular label may sometimes have a very different ancestral history to others who self-identify with that same label. Such "outlier" individuals may complicate interpretation of a labeled population's history by averaging over individuals with different ancestral backgrounds.

Of course, there are numerous other definitions of population that may be useful depending on the question being asked. For example, an unbiased understanding of the genetic variation within a geographic region in the present-day would require a random sample representative of that geographic region, regardless of how many different ancestral populations are represented in such a sample. However, in the following we focus in part on defining populations as groups of genetically homogeneous individuals. We also note that individuals may descend from multiple such populations. Indeed applications of PCA to data from multiple European populations [8, 9] shows a cline of genetic variation whereby sampled individuals more geographically near one another typically are more genetically related. Therefore a restriction to discrete homogeneous populations may sometimes lead to an incomplete understanding of population structure [13]. While this complicates inference, clustering algorithms have been adapted to cope with individuals deriving their ancestry from multiple sources, as we discuss below.

3.2 Clustering based on allele frequency patterns

First, start with the relatively simple scenario where the genome from each of N sampled individuals of ploidy J is derived entirely from one of K genetically distinct populations. At each locus $l \in [1, ..., L]$, let $p_{kl}(a)$ be the frequency of allele type a in population k, with A_l possible allele types at locus l. In the following, for shorthand we will let P represent the set of allele frequencies $\{p_{kl}(a)\}$ for each allele type at all L loci in all K populations. Under a simplified model that assumes random mating and ignores LD and admixture, each allele X_{ilj} for $j \in [1, ..., J]$ (e.g. $J = 2$ in diploids) carried by individual i at locus l is drawn at random according to the allele frequencies at this locus in the population Z_i to which individual i is assigned. I.e.

$$\Pr(X_{ilj} = a | Z_i = k, P) = p_{kl}(a), \tag{1}$$

for each possible allele type a at locus l. Assuming all loci are independent, the probability of observing the full genetic data $X_i \equiv \{X_{i11}, ..., X_{iLJ}\}$ for individual i from population Z_i is:

$$\Pr(X_i | Z_i, P) = \prod_{l=1}^{L} \prod_{j=1}^{J} \prod_{a=1}^{A_l} p_{Z_i l}(a)^{[X_{ilj} == a]}, \tag{2}$$

where $[X_{ilj} == a]$ equals 1 when X_{ilj} is equal to allele type a and is 0 otherwise. Assuming sampled individuals are independent, we have:

$$\Pr(X_1, ..., X_N \mid Z_1, ..., Z_N, P) = \prod_{i=1}^{N} \Pr(X_i \mid Z_i, P) \tag{3}$$

We do not observe P or the Z_is in reality, but instead must infer them using the X_is that we do observe. A natural way to do this is using the Bayesian approach taken by the program STRUCTURE [14], where:

$$\Pr(Z_1, ..., Z_N, \Theta \mid X_1, ..., X_N) \propto \Pr(X_1, ..., X_N \mid Z_1, ..., Z_N, \Theta) \Pr(Z_1, ..., Z_N, \Theta), \tag{4}$$

and the $\{Z_1, ..., Z_N\}$ and Θ are inferred using Markov-Chain-Monte-Carlo (MCMC) [CITE WEGMANN CHAPTER]. In this section $\Theta = \{P\}$, but it will contain additional parameters in the next section. In the STRUCTURE model, the cluster assignments (Z_i) are assumed to be independent across individuals. Therefore

$$\Pr(Z_1, ..., Z_N, \Theta) = \Big[\prod_{i=1}^{N} \Pr(Z_i \mid \Theta) \Big] \Pr(\Theta). \tag{5}$$

In this section, $\Pr(\Theta) = \Pr(P)$, which can be broken down into the product of probabilities across each (assumed independent) locus, i.e.

$$\Pr(P) = \prod_{l=1}^{L} \Pr(\vec{p}_{1l}, ..., \vec{p}_{Kl}), \tag{6}$$

where $\vec{p}_{kl} = \{p_{kl}(1), ..., p_{kl}(A_l)\}$ is the vector of frequencies in population k for each of the A_l allele types at locus l. Following [15], the original STRUCTURE model assumes that the $\vec{p}_{kl}$ are independent across clusters, so that equation (6) is equal to $\prod_{l=1}^{L} \prod_{k=1}^{K} \vec{p}_{kl}$. They further assume that each $\vec{p}_{kl}$ follows a Dirichlet distribution with A_l parameters that can be either fixed or estimated along with the P and Z_i terms. An alternative formulation for equation (6), devised by Falush et al [1] and based in part on work described in [16], models correlations

among clusters' allele frequencies by assuming each cluster's allele frequency has drifted independently from that of an ancestral population common to all clusters. This new formulation is intuitively attractive in that allele frequencies generally are highly correlated across closely related populations in real-life, e.g. across human groups sampled from nearby geographic areas.

Finally, to complete the original formulation of STRUCTURE, each individual is equally likely a priori to be assigned to any of the K clusters, so that:

$$\Pr(Z_i = k \mid \Theta) = 1/K. \tag{7}$$

MCMC is used to sample P and $\{Z_1, ..., Z_N\}$ conditional on the data. This can be accomplished by first proposing initial values for $\{Z_1, ..., Z_N\}$, e.g. by randomly assigning the N individuals to the K clusters with equal probability. Then, at each MCMC iteration, all parameters of P are sampled using the above probabilities conditional on these initial values of $\{Z_1, ..., Z_N\}$. New values for $\{Z_1, ..., Z_N\}$ are then sampled conditional on these updated P parameters, and this iteration between sampling P and sampling $\{Z_1, ..., Z_N\}$ continues until the algorithm converges for these parameters.

The above assumes that K is known. Choosing an appropriate value of K is a notoriously challenging statistical problem [17] that may depend on sampling strategy and other factors. Indeed there is no true value of K, so approaches aim to find a K that scores best according to some intuitive criteria. Several suggestions have been proposed in the literature to choose the "best" K using heuristics [14, 17, 18]. In contrast, the approach taken in STRUCTURAMA [19] models K as a random variable using an approach proposed by Pella and Masuda [20]. However, perhaps the most prevalant usage of STRUCTURE-based clustering algorithms (e.g. [21]) is to cluster using multiple values of K, and then compare the cluster results at each value and attempt to interpret results in light of historical information from other resources such as linguistics, archaeology and anthropology. We discuss limitations of interpreting clustering in Section 3.6 below.

3.3 Incorporating admixture

Many individuals derive from recent mixtures of individuals with genetically different ancestries, a topic we will address further in Section 5. Therefore, assigning each individual to a single cluster can miss important information about recent ancestry. To address this, the original STRUCTURE model allows for each allele at each locus within an individual to have its own ancestry. In particular, we now let $Z_i = \{Z_{i11}, ..., Z_{iLJ}\}$, with Z_{ilj} the cluster to which individual i derives its jth allele at locus l. The Z_is in equations (1) and (2) are replaced with Z_{ilj}, and equation (7) is replaced with:

$$\Pr(Z_{ilj} = k \mid \Theta) = \Pr(Z_{ilj} = k \mid Q) = q_{ik}, \tag{8}$$

where q_{ik} can be thought of as the proportion of DNA for which individual i is most closely related to that of individuals in cluster k, with Q containing the set of all such proportions across all individuals and clusters. In this admixture model, note that $\Theta = \{P, Q\}$ in equations (4) and (5), and that now $\{Z_1, ..., Z_N\}$, P and Q are jointly inferred using MCMC. Here $\Pr(\Theta) = \Pr(P)\Pr(Q)$ in equation (5) and, assuming that the ancestry proportions are independent across individuals, we have that:

$$\Pr(Q) = \prod_{i=1}^{N} \Pr(q_{i1}, ..., q_{iK}). \tag{9}$$

$\Pr(q_{i1}, ..., q_{iK})$ is assumed to be Dirichlet distributed, with K parameters that again can be fixed or inferred using MCMC [14, 1].

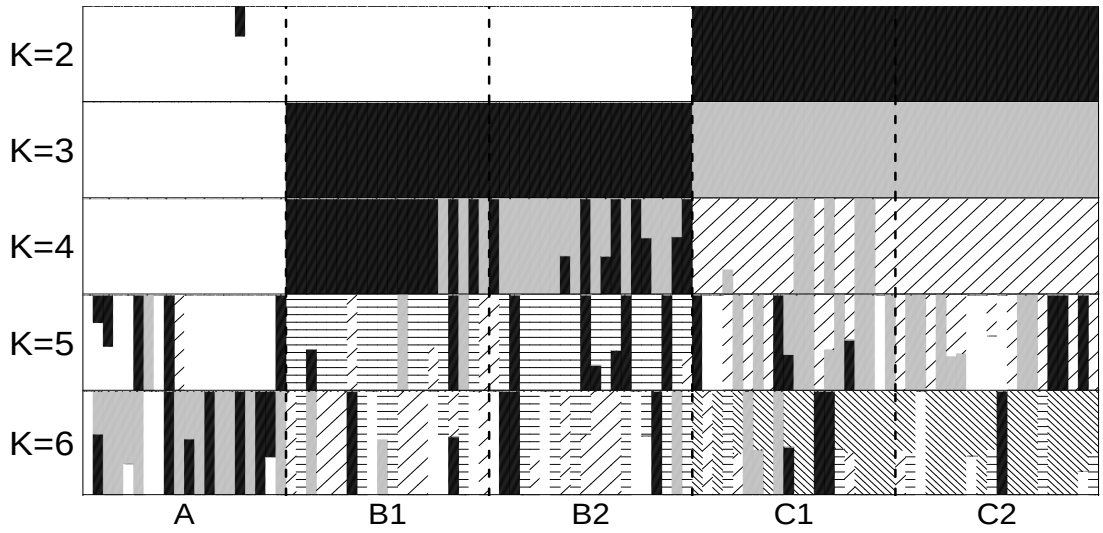


Figure 3: Inferred cluster assignments (different patterns/colors) for each individual (column) for the simulations of populations {A,B1,B2,C1,C2} related as shown in Fig 2a, as inferred by ADMIXTURE for $K = 2$ to 6 clusters.

The models implemented in STRUCTURE can carry a relatively high computational burden due to its MCMC sampling scheme. However, subsequent models that are much more computationally efficient, and thus applicable to larger genetic variation datasets, have been developed that infer P and Q using a variational Bayes framework (fastSTRUCTURE, [18]) or maximum likelihood techniques calculated using Expectation-Maximisation (FRAPPE, [22]) or optimisation techniques (ADMIXTURE, [17]) instead of MCMC. In particular ADMIXTURE and FRAPPE have been used in a large number of studies to explore the ancestry of worldwide groups (e.g. [23, 24, 25, 26, 27, 28, 29]), typically by comparing results across various values of K. Figure 3 shows results from applying ADMIXTURE with K=2-6 to the simulated data illustrated in Fig 2a. After cross-validation [17], the best inferred K out of 2-9 is for $K = 2$, which only separates populations {A,B1,B2} from populations {C1,C2}. This highlights the challenges of inferring the "best" K, as $K = 3$ cleanly separates groups A, B and C, and $K = 4$ shows some ability to separate B1 from B2. This clustering reflects, with some noise, the patterns in the PCA plot of Figure 2b, where visual separation of A, B1, B2 and C={C1,C2} is apparent.

3.4 Incorporating admixture LD

While assuming independence across loci ignores LD, Falush et al [1] introduced an update to STRUCTURE that allows for correlations in cluster membership among adjacent loci. Their model attempts to capture the mosaic-like pattern of DNA expected in an individual that descends from the admixing of ancestors from genetically distinct populations as illustrated in Figure 1. This leads to admixture LD, as an individual's genome consists of a segment of contiguous loci that all share a common ancestral source, followed by another segment from a different ancestral source, with each of these ancestral source groups (ideally) represented by a different cluster. To mimic this, in their formulation the cluster membership at a locus depends on that of the previous locus, forming a Markov chain. In particular, equation (8), which models each locus l independently of other loci, is replaced by a different model reflecting this dependence structure by using the following formulae:

$$\Pr(Z_{i1j} = k|\Theta) = \Pr(Z_{i1j} = k|r, Q) = q_{ik}, \tag{10}$$

and

$$\Pr(Z_{i(l+1)j} = k' \mid Z_{ilj} = k, r, Q] = \begin{cases} \exp(-g_l r) + (1 - \exp(-g_l r))q_{ik} & \text{if } k' = k \\ (1 - \exp(-g_l r))q_{ik'} & \text{otherwise,} \end{cases} \quad (11)$$

where g_l is the genetic distance (in Morgans) between loci l and $l+1$, assumed known, and r can be thought of as a scaling parameter (related to the number of generations ago in which populations admixed) that typically is estimated using the data. Note then that, under this "linkage" model of STRUCTURE, $\Theta = \{P, Q, r\}$ in equations (4) and (5), with r an additional parameter to be estimated from the data. Though the formulation in equations (10) and (11) assumes haploid data, Falush et al [1] describe an approach for dealing with uncertain or unknown phase (see Marchini chapter).

The model defined by equations (10) and (11) is equivalent to assuming that the number of switches in cluster membership for haploid genome j of individual i follows a Poisson distribution with mean $g_l r$ between loci l and $l+1$. Therefore high values of r and/or regions with high rates of recombination (i.e. regions with high g_l) are expected to switch cluster membership between loci more frequently. This formulation mirrors expectations under a simple model of instantaneous admixture between two or more groups (e.g. with each group represented by a distinct cluster) that occurred r generations ago, followed by random mating among individuals from the admixed population (Figure 1). Under this setting, each generation of random mating incurs recombinations that break down the lengths of contiguous DNA segments inherited from each admixing group. As a result, within the DNA of descendents of the admixed population living r generations after the admixture event, the boundaries of DNA segments inherited from each admixing group will form a Poisson process of rate r per Morgan. The top part of equation (11) reflects the probability that either there is no switch between l and $l+1$, which has probability $\exp(-g_l r)$, or there is $\geq$1 switches, which has probability $(1 - \exp(-g_l r))$, with the final switch resulting in a return to the same cluster k with probability q_{ik}. The bottom part of equation (11) gives the probability there is $\geq$1 switch between l and $l+1$, and that the final switch results in a switch to cluster k'.

While this new linkage model of [1] is a more accurate reflection of the process of admixture in organisms that experience homologous recombination, it can be more computationally expensive. Therefore, more efficient algorithms such as ADMIXTURE and FRAPPE that ignore admixture LD are often applied in practice to cope with the large scales typical of current human data collections. However, computational considerations may be less limiting when considering applications to other organisms.

3.5 Incorporating background LD: using haplotypes to improve inference

While the linkage model of STRUCTURE models admixture LD, a limitation is that it does not model the background LD occurring *within* each ancestral population. As a result, STRUCTURE and related models assume that each locus is not linked to other neighboring loci, and typically remove loci until all pairwise combinations of nearby loci are not strongly correlated (e.g. have squared-correlation coefficient r^2 less than some threshold) to meet this assumption. This does not take advantage of recent advances in high thoroughput genotyping technology that allow the routine generation of high density SNP data, including sequencing data. The alternative software fineSTRUCTURE [7] accounts for LD when classifying individuals into genetic clusters. Relative to STRUCTURE/ADMIXTURE/FRAPPE and related approaches, fineSTRUCTURE does not require removing data, and furthermore exhibits increased power in some scenarios as a result of modeling the associations among tightly-linked loci.

To do so, Lawson et al [7] first employ the "chromosome painting" technique CHROMOPAINTER that compares genetic variation patterns in one sampled individual to that in all others. Individuals are assumed phased, and each haploid genome of individual i is compared to the $D = 2(N-1)$ haploid genomes of all $N-1$ other sampled individuals. We will let $Y_{ij} \equiv \{Y_{i1j}, ..., Y_{iLj}\}$ represent the genetic variation data at all L loci for the (phased) jth haploid genome of individual i (where there are $J = 2$ such haploids in humans and other diploids), with $H_d \equiv \{H_{d1}, ..., H_{dL}\}$ analogously representing the genetic variation data at all L loci for the (phased) dth haploid genome that individual i's genome is being compared to. CHROMOPAINTER aims to identify which haploid genome $d \in [1, ..., D]$ among these other sampled individuals is most recently related to the jth haploid of i at each locus. To do so, CHROMOPAINTER uses an approach derived from the copying model of Li & Stephens [30], which attempts to capture key features of coalescent modelling while remaining computationally tractable. Two haploids will have relatively similar DNA sequences if they share a recent ancestor, so intuitively CHROMOPAINTER is attempting to identify which DNA sequence(s) among D has allelic patterns that best match that of the jth haploid of i (see Figure 6a). Therefore, conditional on sharing an inferred most recent ancestor with d at a locus l, CHROMOPAINTER puts high probability on $Y_{ilj} = H_{dl}$. The haploid that is most recently related to the jth haploid of i is expected to switch along the chromosome, because of historical recombination events along the ancestral graph relating the sample. Following Li & Stephens [30], [7] assume these switches occur as a Poisson process, and that the probability of observing Y_{ij} conditional on $H_1, ..., H_D$ follows a Markov chain, with:

$$\Pr(Y_{i1j} = d|\Phi) = q_{id}, \tag{12}$$

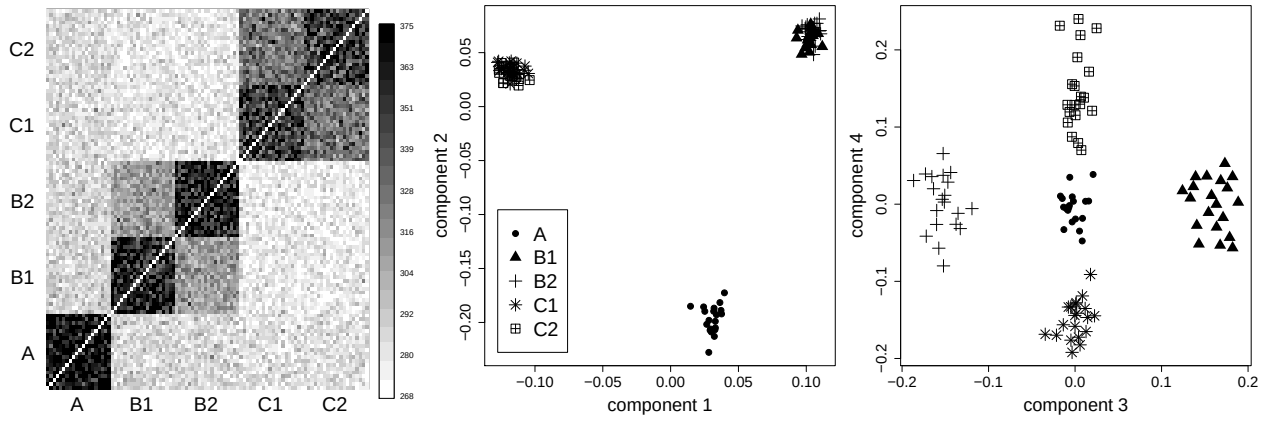
and

$$\Pr(Y_{i(l+1)j} = d' \mid Y_{ilj} = d, \Phi] = \begin{cases} \exp(-g_l\rho) + (1-\exp(-g_l\rho))q_{id} & \text{if } d' = d \\ (1-\exp(-g_l\rho))q_{id'} & \text{otherwise.} \end{cases} \tag{13}$$

Here $\Phi = \{\rho, q_{i1}, ..., q_{iD}\}$ are the parameters of the model, with q_{id} the probability of individual i matching to donor d, which can be fixed or estimated (this is fixed to be $1/D$ in fineSTRUCTURE applications), g_l the genetic distance between loci l and $l+1$ as before, and ρ a scaling constant that is estimated using the data. This is analagous to the linkage STRUCTURE model equations (10) and (11), but a critical difference is that it captures background LD by comparing genetic variation data among sampled individuals. While there is no straight forward interpretation of ρ here, intuitively ρ summarizes the total amount of diversity among the $2N$ haploid genomes. For example, ρ will generally be lower if comparing genetic variation patterns among individuals sampled within Europe relative to comparing genetic variation patterns among individuals sampled world-wide.

Under the Hidden Markov Model (HMM) structure defined by equations (12) and (13), it is straight-forward (e.g. see [31]) to calculate $W_i \equiv \{W_{i1}, ..., W_{iN}\}$, where W_{ih} is the expected number of haplotype segments that the J haploid genomes of individual i match to those of individual $h \in [1, ..., N]$. Here $W_{ii} = 0$, as individual i's haploid genomes are not matched to each other. An example of the matrix of W_{ih} across all pairings of individuals (i, h) is provided in Figure 4a for the simulated individuals from Fig 2a. Lawson et al [7] show that when setting $\rho = \infty$ in equation (13), which therefore models loci as unlinked, the matrix containing all W_i closely relates to the genetic information used by STRUCTURE [14] and by principal components analysis (e.g. using EIGENSTRAT [4]), motivating use of the W_i as summary statistics in practice.

Roughly speaking, the program fineSTRUCTURE then classifies individuals into K clusters based on which share similar W_i vectors. In particular fineSTRUCTURE assumes (up to



(a) W_{ih} of CHROMOPAINTER (b) PCA of W_{ih} matrix

Figure 4: (a) Expected number of haplotype segments (W_{ih}) by which each individual i (column) matches to each other individual h (row) for the simulations of populations {A,B1,B2,C1,C2} related as shown in Fig 2a, as inferred by CHROMOPAINTER. (b) The first four principal components of a PCA of the matrix in (a), with each point an individual from one of the 5 populations, after re-setting the diagonal and scaling as described in [7]. Note that visualising the data using the heatmap in (a) shows clear population structure without requiring working through the principal components.

an adjustment factor c below) that W_i follows a multinomial distribution, with $\frac{B_{(Z_i)k}}{n_{ik}}$ the probability that individual i matches a haplotype segment to an individual from cluster k, where Z_i is the cluster assignment of individual i as before and n_{ik} is the number of individuals from cluster k that individual i's haploid genomes were compared to using equations (12) and (13). Note that $B_{(Z_i)k}$ is therefore the (unknown) total proportion of haplotype segments that an individual from cluster Z_i matches to all individuals from cluster k. Letting $B_{(Z_i)} = \{B_{(Z_i)1}, ..., B_{(Z_i)K}\}$ and $W = \{W_1, ..., W_N\}$, [7] set:

$$\Pr(W \mid Z_1, ...Z_N, B_1, ..., B_K, K) = \prod_{i=1}^{N}\prod_{h=1}^{N}\Big(\frac{B_{(Z_i)(Z_h)}}{n_{i(Z_h)}}\Big)^{W_{ih}/c}, \tag{14}$$

where c is a "likelihood tuning" parameter, inferred using the data, to account for the non-independence of the W_{ih} terms violating the assumption of the multinomial distribution. Each individual is assigned to only a single cluster in fineSTRUCTURE, analagous to the model of STRUCTURE that does not allow for admixture. Each $B_m = \{B_{m1}, ..., B_{mK}\}$ for $m \in [1, ..., K]$ is assumed to follow a Dirichlet distribution with additional parameters estimated using the model, with these parameters, $\{Z_1, ..., Z_N\}$, and K estimated using MCMC steps derived in part from [32] and [20]. FineSTRUCTURE also constructs a tree relating the clusters, by merging pairs of clusters in a greedy fashion until only two remain. This tree can highlight which groups of clusters are most genetically related to one another. However, the authors caution against interpretating the fineSTRUCTURE tree as an ancestral tree relating the groups, as the tree is based on the model and affected by sampling strategy (e.g. sample sizes of different groups) among other factors.

Overall the fineSTRUCTURE model is similar to STRUCTURE, but clusters based on the W_i vectors that capture information about haplotype sharing patterns, rather than clustering based on allele frequencies. Matching haplotype patterns enables an improved inference of shared recent ancestors, which in turn can provide more discriminatory power to distinguish among populations that have only recently become isolated from one another. This is

demonstrated in Fig 4b, which applies PCA to the matrix of W_{ih} in Fig 4a. The first four eigenvectors of this PCA show a clear visual separation of all 5 simulated populations. When applied to these simulated data, fineSTRUCTURE therefore infers the correct number of populations ($K = 5$) and correctly assigns each individual to their true population. In contrast, ADMIXTURE has difficulty distinguishing the DNA of individuals from sub-populations B1 and B2 that split at 2000 years ago, with little resolution to distinguish the DNA of individuals from sub-populations C1 and C2 that split at 1000 years ago (Figure 3). ADMIXTURE results mimic the PCA analysis of Fig 2b, which similarly uses allele frequencies rather than haplotypes and only partly separates individuals from populations C1 and C2. In other words, using haplotypes appears to make the most difference when trying to capture very subtle genetic variation (e.g. distinguishing C1 from C2), while it may not make much a difference when trying to characterise less subtle variation (e.g. distinguishing A from B={B1,B2} from C={C1,C2}).

As a real data example, an application of fineSTRUCTURE to individuals sampled across the United Kingdom elucidated a strong correlation between genetics and geography across England that was not as apparent when analysing the same data using PCA [33]. Therefore, among human groups, using haplotypes might show noticeably greater resolution when considering variation within a country, at least for those with predominantly European ancestry. It is worth noting that other definitions of W_{ih} can also be used in equation (14) or a related probability function, for example replacing the definition of W_{ih} above with an estimate of the total amount of genome shared identical-by-descent (IBD) between individuals i and h as inferred by e.g. [34]. For a comprehensive look at the use of different "similarity" or "coancestry" matrices W, as well as a comparison of different algorithms to cluster individuals using these matrices, see [35].

In addition to having increased resolution, both by not having to remove loci a priori and from using haplotype information, an additional benefit is that analyses based on haplotypes appear to be less affected by genotype chip ascertainment strategies [36, 37]. However, a disadvantage of CHROMOPAINTER/fineSTRUCTURE is that they require the phase of each individual to be estimated a priori. They are also computationally expensive in their current implementation relative to e.g. ADMIXTURE and PCA. Perhaps more importantly, they also currently only allow a model that classifies each individual into a single cluster, rather than the more flexible versions of STRUCTURE that allow individuals to be classified into multiple clusters. In practice, this typically means that individuals with substantial amounts of admixture will be separated into their own clusters rather than cluster with individuals that are closely related to any particular one of the admixing source groups.

3.6 Interpreting genetic clusters

While the motivation for the mathematical models behind these clustering algorithms reflect plausible biological scenarios, they make a number of simplifying assumptions that may be inappropriate in many applications. For example, the choice of K is often arbitrary. Indeed note that fineSTRUCTURE selects K automatically in an attempt to parse individuals into groups of genetically indistinguishable individuals, while STRUCTURE-based approaches often assume fixed smaller values of K (e.g. <10) in order to capture a higher level of genetic differentiation among clusters. However, clusters should not necessarily be interpreted as reflecting K genuine ancestral populations that lived in the past [14]. Therefore, while the STRUCTURE-based clustering models can accurately infer proportions of admixture under certain scenarios, the observation that different individuals are classified to be mixtures of different clusters does not necessarily imply admixture.

In particular individuals that have experienced relatively strong isolation and low effective population size can be assigned to a unique cluster in certain applications with small fixed K.

For example, an application of STRUCTURE with $K = 5$ clusters to Short Tandem Repeat data in individuals sampled from 52 world-wide populations largely separates individuals from major worldwide regions, with clusters roughly corresponding to Africa, Americas, East Asia, West Eurasia, and Oceania [21]. However, at $K = 6$ clusters, primarily only the Kalash, an isolated group from northwest Pakistan, are separated from these world-wide clusters [21]. While the clusters at $K = 5$ likely reflect the relatively long time of isolation between these world-wide groups, the separation of the Kalash at $K = 6$ instead likely reflects isolation of this group from the others on a much more recent time scale [21], highlighting the complications in interpreting the demographic meaning of clusters.

This issue was demonstrated by van Dorp et al [38] in an application to different ethnic and occupational groups from Ethiopia [39]. As explained in more detail by [40], applications of ADMIXTURE to these data at small values of K in three different studies classified a particular labeled group, "Ari Blacksmiths", into their own relatively homogeneous cluster, with individuals from other Ethiopian groups showing various degrees of partial assignment to this cluster [39, 41, 38]. Two of these studies concluded that these observations were consistent with the Ari Blacksmiths reflecting an ancestral source population that subsequently intermixed with the ancestors of some other Ethiopian groups [39, 41]. However, the other study performed additional analyses, including those with techniques described in Section 5, that instead suggested Ari Blacksmiths are recently related to other Ethiopian groups and have similar admixture histories, with the relatively recent isolation and low effective population size of Ari Blacksmiths from other groups likely driving ADMIXTURE inference [38].

In essence these methods, including CHROMOPAINTER heatmaps as in Figure 4a, are descriptions of the data influenced by demographic processes, for which many processes can lead to the same patterns [38, 40]. As with all models discussed here, this suggests caution in over-interpreting the results of clustering algorithms. In the next two sections, we discuss methods that instead attempt to infer specific demographic parameters that lead to genetic variation patterns.

4 Inferring population size changes and split times

In addition to classifying individuals into populations, another major focus of interest is inferring features of populations' demographic histories, including the timings of when populations separated (or became isolated), and the extent and timing of size changes within each population. There are many techniques to assess the demographic history of a population(s), including those that use derivations of F_{ST} and LD (e.g. [42]). Here we will focus on two types of demographic inference models that are prevalent in recent literature: techniques that ignore LD by predicting how different demographic models will affect the observed allele-frequency-spectrum, and techniques that model LD while analysing whole genome sequencing data.

4.1 Allele-frequency-spectrum approaches

One technique towards inferring population demography involves predicting its effect on the allele frequency spectrum (AFS), aka the site frequency spectrum (SFS), which is the distribution of allele counts across loci in a population. The AFS is a complete summary of the data when loci are independent [43, 44], and several researchers have derived expressions for the AFS under a variety of demographic scenarios involving single or multiple populations (e.g. [45, 43, 44, 46]).

As an example, Adams & Hudson [43] consider the demography of a single population. Let m_i be the number of sampled biallelic loci (e.g. SNPs) that have exactly i copies of the derived allele in a sample size of N chromosomes from the population. Given $L = \sum_{i=1}^{N-1} m_i$ total

biallelic unlinked loci have been sampled, then in the absence of ascertainment bias the AFS $(m_1, ..., m_{N-1})$ follows a multinomial distribution:

$$\Pr(m_1, ..., m_{N-1}) = \binom{L}{m_1 \ m_2 \ ... \ m_{N-1}} \prod_{i=1}^{N-1} \gamma_i^{m_i}, \tag{15}$$

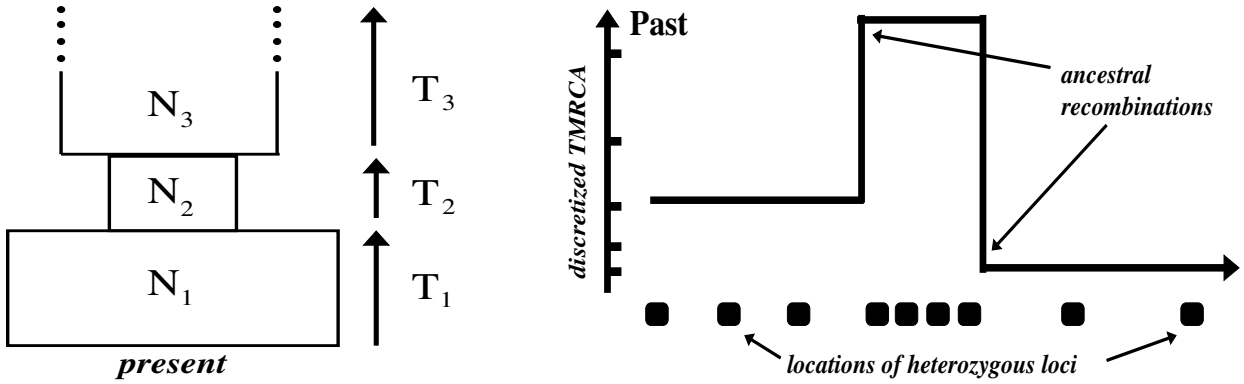
where γ_i is the probability that a locus carries i copies of the derived allele conditional on being polymorphic [47, 48, 43]. Using standard coalescent theory, Adams & Hudson [43] show that the γ_i terms can be predicted using coalescent simulations under a variety of parameters of interest for a particular demographic model, so that maximimum-likelihood estimation can be used to identify the parameters that best fit the AFS. In their application, Adams & Hudson [43] are interested in inferring the time and extent of population size change under a model of exponential population growth following an instantaneous bottleneck, though they note that other demographic models could be considered using equation (15).

Other authors have used different derivations of demographic models that predict the AFS based on the questions of interest, often using similar assumptions though attempting to cope with ascertained SNP data. For example, Marth et al [45] predict the AFS under a scenario where a single population has experienced multiple "epochs" of constant population size, inferring the magnitude of each population size change and the times during which the population remained at each constant size (see Figure 5a). A model that describes population size fluctuations as periods of constant population size over discrete time intervals, with instantaneous population size changes between intervals, is referred to as "piecewise constant population sizes". Keinan et al [49] extended the model of Marth et al [45] to include a second population and predict the split time between the two populations. Meanwhile Gutenkunst et al [44] used different derivations and considered up to three populations. Using linked loci does not affect the expectation of the AFS under these models but can affect the variance, which the above works address by using simulations with linkage [43, 44] or bootstrap re-sampling of linked regions [49] to infer uncertainty around parameters. There are limitations in the amount of information available from the AFS, however, in that very distinct demographic histories can give very similar AFS [50], a familiar drawback common to PCA and clustering algorithms, though the ramifications of this in practice are debated [51].

4.2 Approaches using whole genome sequencing

The increasing availability of sequenced whole genomes of humans [53] has encouraged a new suite of techniques to exploit these data to unearth details of human demography with increased precision. Several methods use approximate coalescent models to fit these sequences and infer parameters of interest. For example, Gronau et al [54] uses a Bayesian, coalescent-based approach that assumes unlinked loci and no recombination. In contrast, many whole-genome sequencing approaches explicitly model linkage between loci [52, 55, 56, 57], thus requiring no a priori reduction of data to unlinked loci.

A widely-used example of the latter is the pairwise sequentially-Markovian coalescent model (PSMC) [52], which is a specialization of the sequentially-Markovian coalescent model of McVean & Cardin [58], described in the McVean Chapter, to two sampled haploid genomes. In particular [52] infers the time to the most recent common ancestor (TMRCA) between the two chromosomes of an individual at each segment of that individual's genome. Assuming a known per nucleotide mutation rate across the genome, the number of mutations within a genomic segment that has not experienced any historical recombination since the TMRCA can be used as a clock to determine the TMRCA within that segment. Assuming a model of piecewise constant population sizes, the effective population sizes leading to the individual's genetic patterns can be inferred over different time intervals in the past. Due to historical recombination,



(a) (b)

Figure 5: (a) Example of the history of a single population divided into epochs (rectangles) of piecewise constant effective population sizes, with an effective population size of N_1 during time period T_1 (leading to present day at bottom), size N_2 during time period T_2, and N_3 during time period T_3. Adapted from [45]. (b) Schematic of the PSMC approach (see main text), illustrating how heterozygous loci (circles) in a diploid individual inform the TMRCA (lines) of the individuals' two sequences, with the TMRCA categorized into discretized piecewise time periods (tick marks on y-axis) and changing along the genome due to ancestral recombinations. Adapted from [52].

the TMRCA changes along the individual's genome, with each such genetic segment in theory providing independent information to assist with inferring this demographic history. Changes in the TMRCA are modeled with a HMM that takes into account linkage between loci, with a transition from a segment with time s to a segment with time t modeled as:

$$\Pr(t \mid s) = (1 - \exp^{-\rho t})\Big[\frac{1}{\lambda(t)} \int_0^{\min(s,t)} \frac{1}{s} \exp^{-\int_u^t \frac{dv}{\lambda(v)}} du\Big] + \exp^{-\rho t}\, \delta(t - s), \tag{16}$$

with ρ the scaled recombination rate, $\lambda(t)$ the population size during the segment with TMRCA t relative to present-day, and $\delta(t - s)$ the Dirac delta function. Conditional on t, the emission probabilities of the HMM determine the probability that each observed locus is homozygous (i.e. no historical mutations since the TMRCA) versus heterozygous, which is modeled as a function of the (unknown) mutation rate and t. The mutation rate, ρ, and the $\lambda(t)$ across discretized time intervals are inferred using an Expectation-Maximisation algorithm.

Intuitively, the PSMC model can be thought of as dividing an individual's genome into regions separated by historical recombinations and counting heterozygotes within each region to infer the TMRCA between the individual's chromosomes in that region (see Figure 5b). Regions with a relatively large number of heterozygotes have a relatively large expected TMRCA. If several regions have their inferred TMRCAs classified into the same discretized time interval, this suggests a smaller effective population size during this time interval. This is because a smaller population size causes an increase in the rate of coalescence events, which is suggested by the large number of regions coalescing during this period.

One attractive feature of PSMC is that inference on population demography can be performed using data from only a single genome, which is of particular interest for analysing DNA from ancient human remains (aDNA) for which relatively few samples have high coverage DNA representing any particular group. However, this can also be seen as a major limitation, in that PSMC has little power to infer population size changes in recent history within the last 20,000 years [55]. In particular, under standard coalescent theory the mean time to TMRCA between two sequences is the effective population size N [59], which among human populations

is inferred to be 2,000-10,000 generations ago (e.g. [60]). In contrast, the expected time to the first coalescence among multiple sequences is more recent, e.g. a factor of $\binom{10}{2} = 45$ times more recent with 10 sequences. In part for this reason, various approaches have been proposed that approximate the coalescent when analysing multiple sequences [55, 56, 57] that have been pre-phased, with the aim of characterizing more recent demography within populations and/or inferring the timing of population splits. In addition, while clean "splits" of populations seem unlikely in reality, such models can also identify groups instead gradually becoming separated with less frequent intermixing over time [55]. However, these multiple individual methods are computationally demanding, at present only allowing joint analysis of tens of samples at a time, and require phased genomes.

As with the AFS, multiple demographic histories can give similar patterns under these approaches, with Mazet et al [61] showing how sub-structure within a population can lead to inaccurate inference of population size changes, suggesting caution is warranted when interpreting results. Furthermore, inferring the time in years of population splits and size changes relies on an estimate of the mutation rate in humans, which is a subject of some debate [62].

5 Identifying/Dating Admixture Events

While the STRUCTURE-based clustering algorithms described in Section 3 can identify whether individuals are admixed and infer the proportions of DNA inherited from each admixing source in some settings, different ancestral histories that either include or exclude admixture can lead to the same patterns in clustering [40]. This makes reliable identification of admixture challenging using these approaches. Correlated-drift approaches [63, 64] [PERHAPS LINK TO LI & AKEY CHAPTER?] can identify admixture and infer proportions with some accuracy, but these approaches can not determine the time period(s) over which admixture has occurred. In this section we describe approaches that attempt to both describe which groups admixed and when they admixed. These approaches often assume that admixture happens in pulses (as exemplified in Figure 1), but can also infer continuous intermixing [65] between sources at a constant rate over a period of time, with some challenges in distinguishing between these scenarios.

As illustrated in Figure 1 and described earlier, assume two or more populations experience an instantaneous admixture event r generations ago. Individuals in the admixed population randomly mate for r generations (i.e. until the present day) following this admixture event. Let a "tract" refer to a segment of DNA that has been inherited intact from a haploid genome in one of the admixing populations without any recombination occurring within the segment. Recall in Section 3.4 that, under some simplifying biological assumptions, such as no crossover interference or impact of gene conversion, the boundaries between tracts of DNA in present-day descendants will follow a Poisson process of rate r per Morgan [1]. On an admixed genome, the probability $\Pr(A \to A; g)$ that two loci separated by distance g (in Morgans) are inherited from the same source A, which has contributed a proportion α of DNA overall to the admixed population, is:

$$\Pr(A \to A; g) = \alpha(\exp^{-gr} + [1 - \exp^{-gr}]\alpha) = \alpha^2 + \alpha(1-\alpha)\exp^{-gr}. \quad (17)$$

This is the probability of either (i) being an intact segment of length g inherited from a single ancestor from population A, which has probability $\alpha \exp^{-gr}$, or (ii) of both loci being inherited from A despite $\geq$1 tract switches between them, which has probability $\alpha^2(1 - \exp^{-gr})$. In the simple case of admixture between only two populations, with the other population B contributing a proportion $\beta = 1 - \alpha$ of DNA overall to the admixed population, the analogous

expression to equation (17) for $\Pr(B \rightarrow B; g)$ simply replaces α with β. The probability $\Pr(A \rightarrow B; g)$ that two loci separated by g are inherited from different populations is:

$$\Pr(A \rightarrow B; g) = \Pr(B \rightarrow A; g) = \alpha\beta(1 - \exp^{-gr}). \quad (18)$$

Analogous expressions can be derived for more complex cases involving >2 admixing populations and >1 pulse of admixture [66].

A key difference between equations (17) and (18) is that the former decays with increasing g, while the latter increases with g, with both having rate gr. This makes it possible in theory to disentangle which sampled groups best reflect source A versus source B, a feature that is exploited by the program GLOBETROTTER [66] described below.

In this section, we describe two different types of approaches to infer the dates and proportions of admixture. The first type explicitly infers local segments of DNA that are inherited from different admixing source groups. The second type considers associations among loci that are attributable to admixture LD, without having to directly identify which DNA segments are inherited from each source group. Figure 6a illustrates the intuition behind the various approaches mentioned in this section. In all methods that date admixture based on using tracts of DNA segments, typically only admixture within the last few hundred generations can be detected reliably, since for older events most tracts will have decreased to lengths not distinguishable from background noise.

5.1 Inferring DNA segments inherited from different sources

The Poisson process for delineating tracts of DNA from admixing sources provides a means of modeling the distribution of lengths of contiguous tracts inherited from each admixing source. Therefore, accurately identifying which DNA segments in the genomes of present-day admixed individuals are derived from each source enables estimation of the proportions of DNA contributed by each source and how many generations ago the admixture occurred.

Typically, statistical models are used to compare the genetic variation data of present-day admixed individuals to that of "surrogate" individuals that are meant to represent genetically each of the admixing source groups, identifying contiguous tracts of DNA inferred to descend from each admixing source as in Figure 6a. The sizes of these inferred tracts are then compared to that expected under different models of migration, to determine e.g. the best-fitting date(s) of admixture and proportions of ancestry inherited from each admixing source [67, 68]. Several different algorithms have been proposed to do this matching (e.g. [69, 70, 71, 25, 72, 73, 74, 75, 76]), with the details of many summarized in [74].

As an example, Figure 6b summarises results from applying RFMix [75], which uses linear discriminant analysis to match segments in admixed individuals to those from reference populations, to simulated data from [66]. Here 20 simulated admixed individuals were composed of tracts of DNA copied intact from individuals randomly chosen from the source populations, with tract sizes sampled according to an exponential distribution as described in [71]. The simulations presented here generated admixed genomes composed of 474,491 autosomal SNPs, each carrying DNA from African and European populations assuming a single instantaneous pulse of admixture between them occurring 30 generations ago. Twenty-one Yoruban individuals from Nigeria were used to represent the African population and 28 individuals from France were used to represent the European population, with 20% of the DNA coming from the French individuals. RFMix was then applied to the admixed genomes using the genomes of 22 Mandenka individuals as surrogates for the African source, and 23 English, Irish, Scottish and Welsh individuals as surrogates for the European source. Figure 6b shows a histogram of the segment sizes matched entirely to the African or French reference groups, as inferred by RFMix, ignoring tracts smaller than 1cM. The dashed line in this figure shows the expected

distribution of tracts, summed across both sources. While equation (17) cannot be used directly for this expectation, since it allows for tracts to be inherited from different sources between the two loci, an expression for $\widetilde{\Pr}(A \to A; g)$, the probability that all tracts between two loci separated by distance g are from A, is:

$$\widetilde{\Pr}(A \to A; g) = \sum_{x=0}^{\infty} \left[\alpha^{x+1}(gr)^x \exp^{-gr} / x!\right]. \tag{19}$$

Equation (19), and its analog $\widetilde{\Pr}(B \to B; g)$, using $r = 30$ and $\alpha = 0.2$, provide the probability distributions to calculate the expected line in Figure 6b. In general there is good agreement between this expectation and the observed distribution inferred by RFMix, highlighting the utility of these approaches that explicitly infer local ancestry tracts.

A limitation of these ancestry tract inference methods is that many, though not all [76, 70], require pre-specified reference individuals to act as surrogates to the admixing sources. Therefore, their accuracy depends on how well these surrogates genetically match the original admixing source groups. An issue for all such approaches is that the admixing sources must be distinguishable using only a few SNPs in a local region. For these reasons, these approaches are typically only applicable in humans to admixture scenarios that involve recent intermixing among different continental groups, as is the case in Latin and African Americans. In such cases, tracts are both easier to identify due to their increased length and the genetic differentiation between admixing sources, and it is more likely that sampled extant populations well reflect the original admixing sources.

5.2 Measuring decay of linkage disequilibrium

Various approaches have been proposed to identify and describe admixture without having to directly infer local tracts of DNA inherited from each admixing source. These approaches instead measure the decay of admixture LD versus genetic distance.

For example, the techniques ROLLOFF [77, 78, 63] and ALDER [79] measure the decay in LD versus genetic distance among pairs of biallelic loci (e.g. SNPs), weighted by each locus's ability to distinguish between the original admixing populations. In particular, again consider an admixed population formed by an instantaneous admixture event between populations A and B occurring r generations ago, with proportion α of the DNA from population A and the rest from B, followed by r generations of random mating (Figure 1). Assuming an infinite population size, the covariance $\mathrm{cov}(x, y)$ between two biallelic loci x and y, which are separated by distance g (in Morgans) and in linkage equilibrium in both A and B, in a diploid individual from this admixed population is:

$$\mathrm{cov}(x, y) = 2\alpha(1-\alpha)(p_{Ax} - p_{Bx})(p_{Ay} - p_{By}) \exp^{-gr}, \tag{20}$$

where p_{Ax} and p_{Bx} are the proportions of a particular allele type in populations A and B, respectively, at locus x [80, 79]. As x and y are in linkage equilibrium in A and B, this covariance between x and y is attributable to the admixture. Only loci with allele frequency differences between A and B contribute to equation (20). Therefore, using surrogates for populations A and B, ROLLOFF and ALDER weight the sample covariance (or correlation) between pairs of biallelic loci in the admixed population by the sample allele frequency differences at these loci in the two surrogate populations. They then fit the decay in this weighted covariance (or correlation) to an exponential function that decays with rate r per Morgan, in order to estimate r. Furthermore, in certain cases the amplitude of these decay curves can be used to infer the proportion of admixture α [79]. The model has also been extended to consider multiple admixture events at different times, which incorporates additional exponential functions to

equation (20), each with rate equal to the number of generations since their respective admixture event [81]. Figure 6c shows the inferred decay in the association between (weighted) pairs of SNPs when applying ROLLOFF to the simulated data described in Section 5.1, using the same surrogate groups. Note how well this decay fits an exponential decay with rate 30, which is the true date of admixture.

The program GLOBETROTTER [66] takes an alternative approach to date and describe admixture. First CHROMOPAINTER matches segments of DNA within (pre-phased) individuals of the putatively admixed population to that of surrogate populations' (pre-phased) individuals. GLOBETROTTER then measures the association among pairs of segments separated by genetic distance g that are matched to surrogate populations A' and B', following the theory defined by equations (17) and (18) and using mixture models to account for biases in the CHROMOPAINTER inference that may be due to e.g. unequal sample size. This approach can increase power by using haplotype information from tightly linked SNPs, rather than allele frequency differences, to discriminate among populations. Fig 6d shows the inferred correlation among segments matched to a particular surrogate population (Mandenka) when applying GLOBETROTTER to the same simulations and surrogates as above. These illustrate a tighter curve around the true date, increasing the precision slightly even in this relatively easy admixture example.

GLOBETROTTER also leverages information on which pairs of surrogate populations A' and B' show exponentially increasing curves as predicted by equation (18), allowing inference of which surrogate populations best represent each admixing source group. For this reason, GLOBETROTTER does not require a priori specification of surrogates for each admixing source group, but instead infers each admixing group as a mixture of an unlimited number of sampled surrogate groups. In contrast, ROLLOFF and ALDER infer a single best surrogate group to represent each source by finding the best model fit out of all possible pairings of available surrogates. GLOBETROTTER can also identify multiple admixture events at different times, as well as admixture among >2 sources occurring at approximately the same time. This comes at an increased computational expense, in addition to the requirement of phasing the data a priori.

In each of ROLLOFF, ALDER and GLOBETROTTER, inferring proportions of admixture from each source typically is more challenging than inferring dates of admixture, with the former often suffering from a lack of identifiability. These models also assume "pulses" of admixture occurring over a short time frame. Continuous admixture occurring over several (perhaps continuous) migrations may be more realistic. In the case of continuous admixture, inferred dates from these approaches typically fall in between the start and end dates of continuous admixture, though may be biased towards the most recent date [79, 66]. Also, as in the methods of Section 5.1, accuracy still depends on how genetically differentiated the admixing sources are, and how well genetic patterns in each source are captured by sampled surrogate individuals. Nonetheless, decay of LD due to admixture can still be usefully modeled by these approaches in far more subtle cases relative to approaches that directly identify tracts in Section 5.1. For example, GLOBETROTTER inferred admixture among European-like source groups dated to over 1000 years ago when applied to genetic variation data from British individuals [33].

By averaging over information across the genome, these techniques should be relatively robust to occasional genotyping or sequencing errors that affect only a limited number of genetic regions. They also appear to be robust to using different human genetic maps in practice [66], despite evidence that recombination hotspot locations can vary among human groups [82]. However, this perhaps is not surprising, as admixture LD tends to extend over megabase scales, for which average recombination rates typically are concordant across continental groups [83]. Nonetheless, disparities in genetic maps may affect inference of older admixture events, for

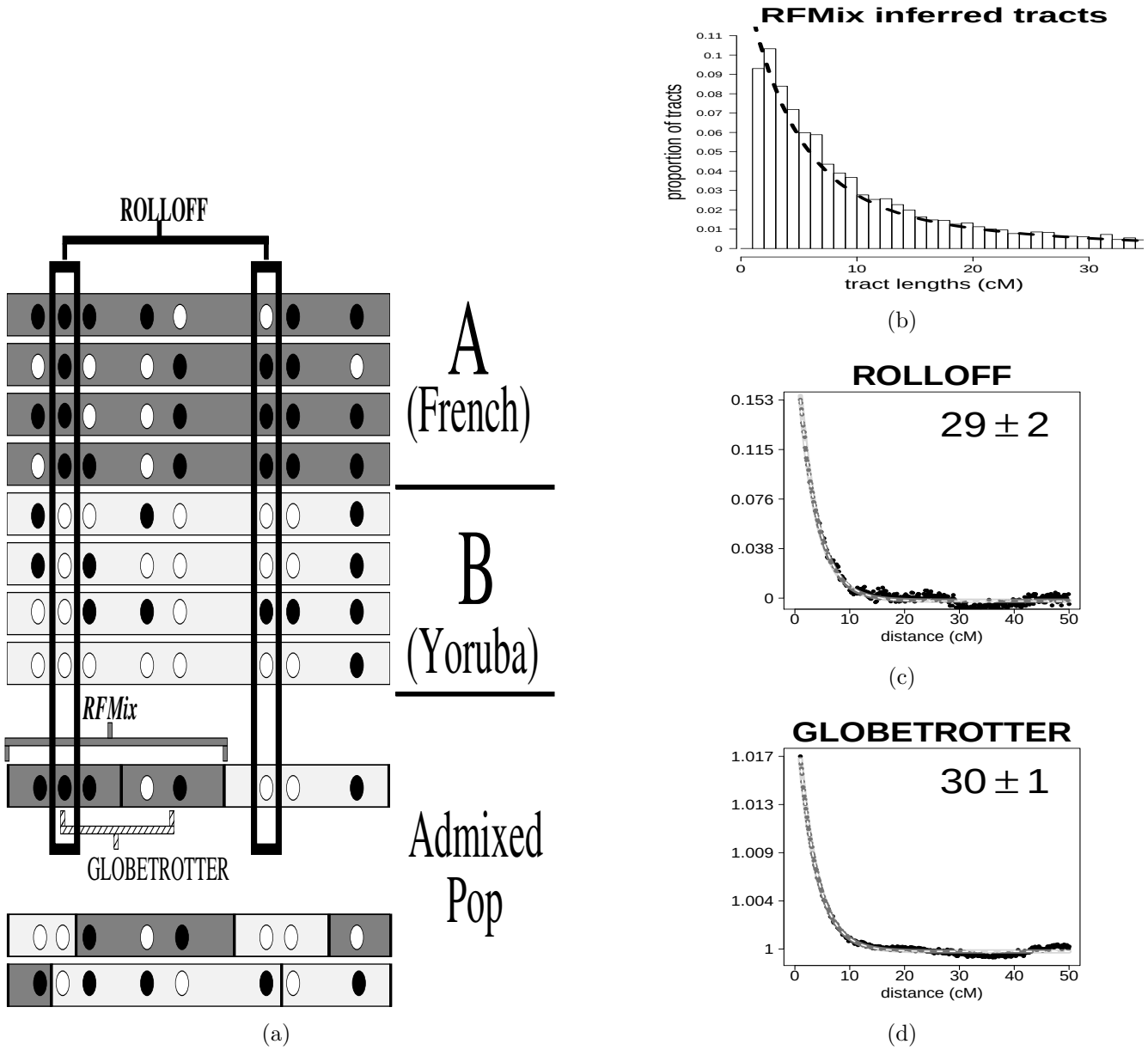


Figure 6: (a) Illustration of approaches to infer dates and proportions of admixture. Two source populations A and B (e.g. French and Yoruba) mix r generations in the past (e.g. $r = 30$) as in Fig 1, generating admixed haplotypes at bottom. Circles denote allele types at each biallelic SNP; dark vertical bars separate segments in admixed chromosomes that exactly match a region of a depicted source haplotype. RFMix [75] attempts to identify contiguous tracts inherited from each of A and B. ROLLOFF/ALDER [63, 79] measure associations among SNPs separated by genetic distance that are informative for distinguishing A and B. GLOBETROTTER [66] measures associations among haplotype segments separated by genetic distance that are inferred by CHROMOPAINTER to match to surrogate haploids for A and B. (b) Distribution of inferred lengths of tracts $\geq$1cM from each source A and B in simulated admixed individuals using RFMix, with the dotted line giving the predicted distribution under the true dates and proportions of admixture. (c)-(d) Inferred association (black lines) between loci versus genetic distance, ignoring loci separated by <1cM, for ROLLOFF and GLOBETROTTER, respectively, for the same simulations, with the true LD decay curve in light grey and the curve for the best fitting model for each method in dashed white. The legend at top right gives the inferred date and standard error, with the latter determined by jackknifing over chromosomes for ROLLOFF and 100 bootstrap re-samples for GLOBETROTTER.

which a large proportion of admixture LD may decay within e.g. a megabase.

6 Conclusion

While this chapter attempts to summarize the models and concepts behind presently popular algorithms for exploring large-scale autosomal data, we remind the reader that this is by no means an exhaustive look at the numerous exciting methods to explore these (and other) questions about demography. Each method presented here has limitations, in terms of both modeling assumptions and computational tractability. Not all limitations and assumptions may be listed here. We refer the reader to the cited literature for more details on particular approaches, though note that exploration of how several of these methods behave in practice are on-going (e.g. see [84] for a discussion on limitations of admixture date inference approaches).

While the approaches outlined here are current state-of-the-art in the field of population genetics, they all still assume a major over-simplification of history. For example, the approaches described above focus on inferring one or two features of demographic history, such as population substructure or admixture or population size changes, while ignoring the other factors that may have altered observed genetic variation patterns. In contrast, there are also more comprehensive approaches that try to infer jointly the ancestral relationships among dozens of sampled populations, including the order of splits and subsequent drift effects (e.g. related to population size changes), as well as admixture events among the ancestors of these populations [64, 85]. However, exhaustively exploring all possible scenarios of splits, admixture and population trees is intractable, so that simplifying assumptions (such as assuming a fixed tree topology) must be used in practice to reduce the search space. For this reason, techniques in this chapter that simplify the problem by e.g. inferring only one or two of these demographic processes will likely remain important and shed light on the demographic history of humans in a piecemeal fashion.

Furthermore, increasingly larger datasets are becoming available for human populations, from massive whole genome sequencing studies of different populations such as UK Biobank (http://www.ukbiobank.ac.uk/) and China Kadoorie Biobank
(http://www.ckbiobank.org/site/) and potentially through large databases of customers' genotypes acquired through genetic ancestry testing companies [86]. These large data resources will enable better understanding of demography, for example allowing identification of individuals that share ancestors more recently than is typically seen is collections of smaller sample size, but will require substantial computational improvements to fully extract the rich available information. In addition to the rapidly increasing resources from present-day populations, techniques to reliably extract high quality DNA from ancient human remains (aDNA) are facilitating the increasing availability of genetic resources from numerous historical cultures and time periods (e.g. [87]). These aDNA samples are already proving extremely valuable for our understanding of ancient human history and will continue to do so. Overall these forthcoming data resources will increase the resolution of genetic studies for unearthing details of human history, suggesting that we are just scratching the surface with current applications.

Acknowledgements

I thank David Balding and Mark Beaumont for their helpful comments that improved the chapter considerably. GH is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 098386/Z/12/Z) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

References

[1] D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[2] J.K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 2001.

[3] Menozzi, P. and Piazza, A. and Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science*, 201:786–792, 1978.

[4] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:504–509, 2006.

[5] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Gen*, 2(12):e190, 2006.

[6] K.J. Galinsky, G. Bhatia, P.R. Loh, S. Georgiev, S. Mukherjee, N.J. Patterson, and A.L. Price. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet*, 98(3):456–72, 2016.

[7] D.J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, 2012.

[8] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergman, M.R. Nelson, M. Stephens, and C.D. Bustamante. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.

[9] Lao, O. and Lu, T.T. and Nothnagel, M. and Junge, O. and Freitag-Wolf, S. and Caliebe, A. and Balascakova, M. and Bertranpetit, J. and Bindoff, L.A. and Comas, D. and Holmlund, G. and Kouvatsi, A. and Macek, M. and Mollet, I. and Parson, W. and Palo, J. and Ploski, R. and Sajantila, A. and Tagliabracci, A. and Gether, U. and Werge, T. and Rivadeneira, F. and Hofman, A. and Uitterlinden, A.G. and Gieger, C. and Wichmann, H.E. and Ruther, A. and Schreiber, S. and Becker, C. and Nurnberg, P. and Nelson, M.R. and Krawczak, M. and Kayser, M. Correlation between genetic and geographic structure in Europe. *Current Biology*, 18(16):1241–1248, 2008.

[10] G. McVean. A genealogical interpretation of principal components. *PLoS Genet*, 5(10):e1000686, 2009.

[11] Novembre, J. and Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40:646–649, 2008.

[12] M. Jakobsson, S.W. Scholz, P. Scheet, R.J. Gibbs, J.M. Vanlier, H.C. Fung, Z.A. Szpiech, J.H. Degnan, K. Wang, R. Guerreiro, J.M. Bras, J.C. Schymick, D.G. Hernandez, B.J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H.M. Cann, J.A. Hardy, N.A. Rosenberg, and A.B. Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451:998–1003, 2008.

[13] N.A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J.K. Pritchard, and M.W. Feldman. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1(6):e70, 2005.

[14] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotypes data. *Genetics*, 155:945–959, 2000.

[15] D.J. Balding and R.A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12, 1995.

[16] G. Nicholson, A.V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson, and P. Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:695–715, 2002.

[17] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[18] Raj, A. and Stephens, M. and Pritchard, J.K. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, 197(2):573–589, 2014.

[19] Huelsenbeck, J.P. and Andolfatto, P. and Huelsenbeck, E.T. Structurama: Bayesian inference of population structure. *Evol Bioinform Online*, 7:55–59, 2011.

[20] Pella, J. and Masuda, M. The Gibbs and splitmerge sampler for population mixture analysis from genetic data with incomplete baselines. *Can. J. Fish. Aquat. Sci*, 63:576–596, 2006.

[21] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298:2981–2985, 2002.

[22] H. Tang, J. Peng, P. Wang, and N. Risch. Estimation of Individual Admixture: Analytical and Study Design Considerations. *Genetic Epidemiology*, 28:289–301, 2005.

[23] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.

[24] D.M. Behar, B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, J. Parik, S. Rootsi, G. Chaubey, I. Kutuev, G. Yudkovsky, E.K. Khusnutdinova, O. Balanovsky, O. Semino, L. Pereira, D. Comas, D. Gurwitz, B. Bonne-Tamir, T. Parfitt, M.F. Hammer, K. Skorecki, and R. Villems. The genome-wide structure of the Jewish people. *Nature*, 466:238–242, 2010.

[25] Bryc, K. and Auton, A. and Nelson, M.R. and Oksenberg, J. R. and Hauser, S.L. and Williams, S. and Froment, A. and Jean-Marie Bodo, J.M. and Charles Wambebe, C. and Tishkoff, S.A. and Bustamante, C.D. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA*, 107(2):786–791, 2010.

[26] Metspalu, M. and Romero, I.G. and Yunusbayev, B. and Chaubey, G. and Mallick, C.B. and Hudjashov, G. and Nelis, M. and Magi, R. and Metspalu, E. and Remm, M. and Pitchappan, R. and Singh, L. and Thangaraj, K and Villems, R. and Kivisild, T. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet*, 89:731–744, 2011.

[27] Schlebusch, C.M. and Skoglund, P. and Sjodin, P. and Gattepaille, L.M. and Hernandez, D. and Jay, F. and Li, S. and De Jongh, M. and Singleton, A. and Blum, M.G. and Soodyall, H. and Jakobsson, M. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, 338:374–379, 2012.

[28] Rasmussen, M. and Anzick, S.L. and Waters, M.R. and Skoglund, P. and DeGiorgio, M. and Stafford Jr, T.W., and Rasmussen, S. and Moltke, I. and Albrechtsen, A. and Doyle, S.M. and Poznik, G.D. and Gudmundsdottir, V. and Yadav, R. and Malaspinas, A.S. and White 5th, S.S. and Allentoft, M.E. and Cornejo, O.E. and Tambets, K. and Eriksson, A. and Heintzman, P.D. and Karmin, M. and Korneliussen, T.S. and Meltzer, D.J. and Pierre, T.L. and Stenderup, J. and Saag, L. and Warmuth, V.M. and Lopes, M.C. and Malhi, R.S. and Brunak, S. and Sicheritz-Ponten, T. and Barnes, I. and Collins, M. and Orlando, L. and Balloux, F. and Manica, A. and Gupta, R. and Metspalu, M. and Bustamante, C.D. and Jakobsson, M. and Nielsen, R. and Willerslev, E. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506:225–229, 2014.

[29] Jones, E.R. and Gonzalez-Fortes, G. and Connell, S. and Siska, V. and Eriksson, A. and Martiniano, R. and McLaughlin, R.L. and Gallego Llorente, M. and Cassidy, L.M. and Gamba, C. and Meshveliani, T. and Bar-Yosef, O. and Muller, W. and Belfer-Cohen, A. and Matskevich, Z. and Jakeli, N. and Higham, T.F. and Currat, M. and Lordkipanidze, D. and Hofreiter, M. and Manica, A. and Pinhasi, R. and Bradley, D.G. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun*, 6:8912, 2015.

[30] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–33, 2003.

[31] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 77:257–286, 1989.

[32] Huelsenbeck, J.P. and Andolfatto, P. Inference of Population Structure Under a Dirichlet Process Model. *Genetics*, 175:1787–1802, 2007.

[33] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E.C. Royrvik, B. Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, D.J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, and W. Bodmer. The fine scale genetic structure of the British population. *Nature*, 519:309–314, 2015.

[34] B.L. Browning and S.R. Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–71, 2013.

[35] Lawson, D.J. and Falush, D. Population identification using genetic data. *Annu Rev Genomics Hum Genet*, 13:337–61, 2012.

[36] D.F. Conrad, M. Jakobsson, G. Coop, X. Wen, J.D. Wall, N.A. Rosenberg, and J.K. Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, 38(11):1251–1260, 2006.

[37] G. Hellenthal, A. Auton, and D. Falush. Inferring human colonization history using a copying model. *PLoS Genetics*, 4(5):e1000078, 2008.

[38] L. van Dorp, D. Balding, S. Myers, L. Pagani, C. Tyler-Smith, E. Bekele, A. Tarekegn, M.G. Thomas, N. Bradman, and G. Hellenthal. Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS Genetics*, 11(8):e1005397, 2015.

[39] L. Pagani, T. Kivisild, A. Tarekegn, R. Ekong, C. Plaster, I. Gallego-Romero, Q. Ayub, S.Q. Mehdi, M.G. Thomas, D. Luiselli, E. Bekele, N. Bradman, D.J. Balding, and C. Tyler-Smith. Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. *The American Journal of Human Genetics*, 91(1):83–96, 2012.

[40] Lawson, D. and van Dorp, L. and Falush, D. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. page https://www.biorxiv.org/content/early/2017/11/27/066431, 2017.

[41] J.A. Hodgson, C.J. Mulligan, A. Al-Meeri, and R.L. Raaum. Early Back-to-Africa Migration in the Horn of Africa. *PLoS Genet*, 10(6):e1004393, 2014.

[42] McEvoy, B.P. and Powell, J.E. and Goddard, M.E. and Visscher, P.M. Human population disperal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Research*, 21(6):821–829, 2011.

[43] Adams, A.M. and Hudson, R.R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked Single-Nucleotide Polymorphisms. *Genetics*, 168:1699–1712, 2004.

[44] R.N. Gutenkunst, R. Hernandez, S.H. Williamson, and C.D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, 5(10):e1000695, 2009.

[45] G.T. Marth, E. Czabarka, J. Murvai, and S.T. Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166:351–372, 2004.

[46] Excoffier, L. and Dupanloup, I. and Huerta-Sanchez, E. and Sousa, V.C. and Foll, M. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10):e1003905, 2013.

[47] Wooding, S. and Rogers, A. The matrix coalescent and an application to human Single-Nucleotide Polymorphisms. *Genetics*, 161:1641–1650, 2002.

[48] Polanski, A. and Kimmel, M. New explicit expressions for relative frequencies of Single-Nucleotide Polymorphisms with application to statistical inference on population growth. *Genetics*, 165:427–436, 2003.

[49] A. Keinan, J.C. Mullikin, N. Patterson, and D. Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet*, 39:1251–1255, 2007.

[50] Myers, S. and Fefferman, C. and Patterson, N. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.*, 73:342–348, 2008.

[51] Bhaskar, A. and Song, Y.S. Descartes Rule of Signs and the Identifiability of Population Demographic Models from Genomic Variation Data. *Ann Stat*, 42(6):2469–2493, 2014.

[52] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496, 2011.

[53] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):55–65, 2012.

[54] I. Gronau, M.J. Hubisz, B. Gulko, C.G. Danko, and A. Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet*, 43:1031–1034, 2011.

[55] Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46:919–925, 2014.

[56] Sheehan, S. and Harris, K. and Song, Y.S. Estimating variable effective population sizes from multiple genomes: a sequentally Markov conditional sampling distribution approach. *Genetics*, 194:647–662, 2013.

[57] M. Raghavan, M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M.C. Ávila-Arcos, A.S. Malaspinas, A. Eriksson, I. Moltke, M. Metspalu, J.R. Homburger, J. Wall, O.E. Cornejo, J.V. Moreno-Mayar, T.S. Korneliussen, T. Pierre, M. Rasmussen, P.F. Campos, P.D.B. Damgaard, M.E. Allentoft, J. Lindo, E. Metspalu, R. Rodríguez-Varela, J. Mansilla, C. Henrickson, A. Seguin-Orlando, H. Malmström, T. Stafford, S.S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergström, Y. Xue, V. Warmuth, A.D. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Leboreiro, J.L. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L.G. Davis, E. Heyer, C.P.E. Zollikofer, M.S. Ponce de León, C.I. Smith, V. Grimes, K.A. Pike, M. Deal, B.T. Fuller, B. Arriaza, V. Standen, M.F. Luz, F. Ricaut, N. Guidon, L. Osipova, M.I. Voevoda, O.L. Posukh, O. Balanovsky, M. Lavryashina, Y. Bogunov, E. Khusnutdinova, M. Gubina, E. Balanovska, S. Fedorova, S. Litvinov, B. Malyarchuk, M. Derenko, M.J. Mosher, D. Archer, J. Cybulski, B. Petzelt, J. Mitchell, R. Worl, P.J. Norman, P. Parham, B.M. Kemp, T. Kivisild, C. Tyler-Smith, M.S. Sandhu, M. Crawford, R. Villems, D.G. Smith, M.R. Waters, T. Goebel, J.R. Johnson, R.S. Malhi, M. Jakobsson, D.J. Meltzer, A. Manica, R. Durbin, C.D. Bustamante, Y.S. Song, R. Nielsen, and E. Willerslev. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, 349:841, 2015.

[58] G.A.T. McVean and N.J. Cardin. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B*, 360:1387–1393, 2005.

[59] Hudson, R.R. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, 7:1–44, 1991.

[60] Tenesa, A. and Navarro, P. and Hayes, B.J. and Duffy, D.L. and Clarke, G.M. and Goddard, M.E. and Visscher, P.M. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520–526, 2007.

[61] Mazet, O. and Rodriguez, W. and Grusea, S. and Boitard, S. and Chikhi, L. On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? *Heredity*, 116:362–371, 2016.

[62] Mazet, O. and Rodriguez, W. and Chikhi, L. Revising the human mutation rate: implications for understanding human variation evolution. *Nat Rev Genet*, 13:745–753, 2012.

[63] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient Admixture in Human History. *Genetics*, 192(3):1065–1093, 2012.

[64] J.K. Pickrell and J.K. Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*, 8:e1002967, 2012.

[65] B.M. Henn, L.R. Botigué, S. Gravel, W. Wang, A. Brisbin, J.K. Byrnes, K. Fadhlaoui-Zid, P.A. Zalloua, A. Moreno-Estrada, J. Bertranpetit, C.D. Bustamante, and D. Comas. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet*, 8(1):e1002397, 2012.

[66] G. Hellenthal, G.B.J. Busby, G. Band, J.F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343:747–751, 2014.

[67] J. E. Pool and R. Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181:711–719, 2009.

[68] Gravel, G. Population genetics models of local ancestry. *Genetics*, 191:607–619, 2012.

[69] H. Tang, M. Coram, P. Wang, X. Xhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individ- uals. *Am J Hum Genet*, 79:1–12, 2006.

[70] S. Sriram Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating Local Ancestry in Admixed Populations. *American Journal of Human Genetics*, 82(2):290–303, 2008.

[71] A.L. Price, A. Tandon, N. Patterson, K.C. Barnes, N. Rafaels, I. Ruczinski, T.H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6):e1000519, 2009.

[72] Baran, Y. and Pasaniuc, B. and Sankararaman, S. and Torgerson, D.G. and Gignoux, C. and Eng, C. and Rodriguez-Cintron, W. and Chapela, R. and Ford, J.G. and Avila, P.C. and Rodriguez-Santana, J. and Burchard, E.G. and Halperin, E. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10):1359–67, 2012.

[73] Brisbin, A. and Bryc, K. and Byrnes, J. and Zakharia, F. and Omberg, L. and Degenhardt, J. and Reyonlds, A. and Ostrer, H. and Mezey, J.G. and Bustamante, C.D. PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol*, 84(4):343–364, 2012.

[74] Churchhouse, C. and Marchini, J. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology*, 37(1):1–12, 2013.

[75] Maples, B.K. and Gravel, S. and Kenny, E.E. and Bustamante, C.D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*, 93(2):278–288, 2013.

[76] Y. Guan. Detecting structure of haplotypes and local ancestry. *Genetics*, 196:625–642, 2014.

[77] P. Moorjani, N. Patterson, J.N. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A.L. Price, and D. Reich. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics*, 7(4):e1001373, 2011.

[78] P. Moorjani, N. Patterson, P.R. Loh, M. Lipson, P. Kisfali, B.I. Melegh, M. Bonin, L. Kadasi, O. Riess, B. Berger, D. Reich, and B. Melegh. Reconstructing Roma history from genome-wide data. *PLoS ONE*, 8(3):e58633, 2013.

[79] P.R. Loh, M. Lipson, N. Patterson, P. Moorjani, J.K. Pickrell, D. Reich, and B. Berger. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4):1233–1254, 2013.

[80] R. Chakraborty and K. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Adad. Sci. USA*, 85(23):9119–9123, 1988.

[81] J.K. Pickrell, N. Patterson, P.R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, and D. Reich. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci USA*, 111(7):2632–7, 2014.

[82] D.C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M.J. Rieder, D.A. Nickerson, and M. Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet*, 36(7):700–6, 2004.

[83] A.G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C.D. Palmer, G.K. Chen, K. Wang, S.G. Buxbaum, E.L. Akylbekova, M.C. Aldrich, C.B. Ambrosone, C. Amos, E.V. Bandera, S.I. Berndt, L. Bernstein, W.J. Blot, C.H. Bock, E. Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L.A. Cupples, S.L. Deming, W.R. Diver, J. Divers, M. Fornage, E.M. Gillanders, J. Glessner, C.C. Harris, J.J. Hu, S.A. Ingles, W. Isaacs, E.M. John, W. H. Kao, B. Keating, R.A. Kittles, L.N. Kolonel, E. Larkin, L. Le Marchand, L.H. McNeill, R.C. Millikan, A. Murphy, S. Musani, C. Neslund-Dudas, S. Nyante, G.J. Papanicolaou, M.F. Press, B.M. Psaty, A.P. Reiner, S.S. Rich, J.L. Rodriguez-Gil, J.I. Rotter, B.A. Rybicki, A.G. Schwartz, L.B. Signorello, M. Spitz, S.S. Strom, M.J. Thun, M.A. Tucker, Z. Wang, J.K. Wiencke, J.S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K.A. Zanetti, W. Zheng, R.G. Ziegler, X. Zhu, S. Redline, J.N. Hirschhorn, B.E. Henderson, H.A. Taylor Jr, A.L. Price, H. Hakonarson, S.J. Chanock, C.A. Haiman, J.G. Wilson, D. Reich, and S.R. Myers. The landscape of recombination in African Americans. *Nature*, 476:170–175, 2011.

[84] M. Liang and R. Nielsen. The lengths of admixture tracts. *Genetics*, 197:953–967, 2014.

[85] Lipson, M. and Loh, P.R. and Levin, A. and Reich, D. and Patterson, N. and Berger, B. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Mol Biol Evol*, 30(8):1788–1802, 2013.

[86] Bryc, K. and Durand, E.Y. and Macpherson, J.M. and Reich, D. and Mountain, J.L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *Amer J Hum Genet*, 96:37–53, 2015.

[87] I. Lazaridis, D. Nadel, G. Rollefson, D.C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.